Tight Bounds on Vertex Connectivity Under Sampling

Keren Censor-Hillel, Mohsen Ghaffari, George Giakkoupis, Bernhard Haeupler, Fabian Kuhn

Tight Bounds on Vertex Connectivity Under Sampling

Keren Censor-Hillel* Mohsen Ghaffari† George Giakkoupis‡ Bernhard Haeupler§
Fabian Kuhn¶

Abstract

A fundamental result by Karger [10] states that for any λ-edge-connected graph with n nodes, independently sampling each edge with probability $p = \Omega(\log(n)/\lambda)$ results in a graph that has edge connectivity $\Omega(\lambda p)$, with high probability. This paper proves the analogous result for vertex connectivity, when either vertices or edges are sampled. We show that for any k-vertex-connected graph G with n nodes, if each node is independently sampled with probability $p = \Omega(\sqrt{\log(n)/k})$, then the subgraph induced by the sampled nodes has vertex connectivity $\Omega(kp^2)$, with high probability. If edges are sampled with probability $p = \Omega(\log(n)/k)$ then the sampled subgraph has vertex connectivity $\Omega(kp)$, with high probability. Both bounds are existentially optimal.

1 Introduction

Consider a random process where given a base graph G, each edge or node of G is sampled with some probability p. Given such a random graph process, it is interesting to see how various global connectivity properties of the graph induced by the sampled edges or nodes change as a function of the sampling probability p. If G is the complete n-node graph, sampling each edge independently with probability p results in the classic Erdős-Rényi random graph $G_{n,p}$, for which exact thresholds for the formation of a giant component, global connectivity, and many other properties have been studied (e.g., in [3]). Thresholds for the formation of a giant component are further studied more generally in percolation theory [4]—mostly for graphs G defined by some regular or random lattice. In the context of percolation theory, edge sampling is called bond percolation whereas vertex sampling is referred to as site percolation.

In the present work, we are interested in how the vertex connectivity of a general graph G changes under uniform random vertex or edge sampling. For edge connectivity and edge sampling, the analogous question has been resolved two decades ago. Karger's seminal result [10] showed that for any λ-edge-connected graph with n vertices, sampling edges independently at random with probability $p = \Omega(\log(n)/\lambda)$ results in an $\Omega(\lambda p)$-edge-connected subgraph, with high probability[1]. This was a strong extension of an earlier result by Lomonosov and Polesskii [13], which stated that sampling each edge with probability $\Theta(\log(n)/\lambda)$ leads to a connected subgraph, w.h.p. These sampling results and their extensions were cornerstone tools for addressing various important problems such as various min-cut problems [9, 10], constructing cut-preserving graph sparsifiers [2, 15], max-flow problems [9, 12], and network reliability estimations [11].

As in the case of edge connectivity, studying the vertex connectivity of the subgraph obtained by independently sampling vertices or edges of a k-vertex-connected graph is of fundamental interest. However, the

*Technion, Israel, ckeren@cs.technion.ac.il. Supported in part by ISF grant 1696/14 and by BSF grant 2015803.
†MIT, USA, ghaffari@mit.edu
‡INRIA, France, george.giakkoupis@inria.fr
§Carnegie Mellon U., USA, haeupler@cs.cmu.edu. Research supported in part by NSF grants CCF-1618280 and CCF-1527110.
¶U. of Freiburg, Germany, kuhn@cs.uni-freiburg.de

[1]We use the phrase '*with high probability*' (*w.h.p.*) to indicate that some event has a probability of at least $1 - n^{-\Theta(1)}$.

vertex connectivity case has been recognized as being much harder and progress has been scarce. Until recently, it was not even known whether a $\Theta(n)$-vertex-connected graph stays (simply) connected when nodes are sampled with probability $p = 1/2$. Recently, Censor-Hillel et al. [6] showed that a vertex-sampling probability of $p = \Omega(\log(n)/\sqrt{k})$ is a sufficient condition for connectivity (w.h.p.), and moreover, that the remaining vertex connectivity of the sampled subgraph is $\Omega(kp^2/\log^3 n)$, w.h.p. It remained open whether these two bounds are optimal and what the corresponding bounds for edge sampling look like.

In this paper, we answer these questions and provide tight bounds for the threshold probability for vertex connectivity and the remaining connectivity under both vertex and edge sampling. For a graph $G = (V, E)$ and a vertex set $S \subseteq V$, we denote by $G[S]$ the subgraph of G induced by S.

Theorem 1. *Let $G = (V, E)$ be a k-vertex-connected n-node graph, and let S be a randomly sampled subset of V where each node $v \in V$ is included in S independently with probability $p \geq \alpha\sqrt{\log(n)/k}$, for a sufficiently large constant α. Then the set S is a dominating set and the induced subgraph $G[S]$ has vertex connectivity $\Omega(kp^2)$, with probability $1 - e^{-\Omega(kp^2)}$.*

Theorem 1 improves over [6] in two ways. First, it improves over [6, Theorem 1.7], which only proves simple connectivity for a sampling probability $p = \Omega(\log(n)/\sqrt{k})$, whereas Theorem 1 guarantees connectivity for $p = \Omega(\sqrt{\log(n)/k})$. Second, it improves over [6, Theorem 1.4], which proves a remaining vertex connectivity of $\Omega(kp^2/\log^3 n)$, instead of $\Omega(kp^2)$.

The following is our result about the vertex connectivity after sampling edges of a given graph. To the best of our knowledge, no non-trivial result was known prior to this work.

Theorem 2. *Let $G = (V, E)$ be a k-vertex-connected n-node graph, and let E' be a randomly sampled subset of E where each edge $e \in E$ is included in E' independently with probability $p \geq \alpha \log(n)/k$, for a sufficiently large constant α. Then the graph $G' = (V, E')$ has vertex connectivity $\Omega(kp)$, with probability $1 - e^{-\Omega(kp)}$.*

In the rest of this section, we first give a brief explanation of why the standard techniques used for the edge connectivity case do not work for vertex connectivity, and we present a brief explanation of our approach and how it compares with that of [6]. Then we discuss a simple graph construction that shows the optimality of the bounds in Theorems 1 and 2, and finally, we state some implications of these results.

1.1 Overview of Analysis

The Challenge. To understand the challenge, we briefly explain why tools with a similar flavor to the ones used for edge connectivity do not take us far in the vertex connectivity case. The key to most results about edge sampling is the "cut counting" argument introduced in [8], where it is shown that in a graph of edge connectivity λ, the number of cuts of size at most $\alpha\lambda$ is at most $O(n^{2\alpha})$. Combined with a standard Chernoff argument and a union bound over all cuts, this shows that when independently sampling each edge with probability $p = \Omega(\log(n)/\lambda)$, it holds w.h.p. for the subgraph induced by the sampled edges, that the size of each cut does not deviate from its expectation by more than a constant factor [10]. Hence, in particular, the edge connectivity of the sampled subgraph is $\Omega(\lambda p)$, w.h.p. Unfortunately, the same approach cannot work for vertex connectivity under vertex or edge sampling, because in graphs with vertex connectivity k, even the number of minimum vertex cuts can be as large as $\Theta(2^k(n/k)^2)$ [7].

The Old Approach. In [6], the bound on the sampling threshold for (simple) connectivity is obtained by essentially considering the vertex sampling as a gradual process that happens in phases, and by analyzing the growth of the connected components throughout this process. More precisely, it is shown that when starting from a dominating set, if each node is sampled with probability $1/\sqrt{k}$, then in expectation, the number

of connected components drops by a constant factor.[2] Hence, after $O(\log n)$ phases where in each phase nodes are sampled independently with probability $1/\sqrt{k}$, and thus after an overall sampling probability of $O(\log(n)/\sqrt{k})$, the subgraph induced by the sampled nodes is connected, w.h.p.

This gradual process is not sufficient on its own for proving values of vertex connectivity higher than one. To prove higher remaining vertex connectivity while trying to avoid explicitly working on all cuts, [6] developed the notion of connected dominating set (CDS) packings[3]. This notion serves as a certificate for large vertex connectivity (among other applications). Particularly, it is shown that after sampling with probability p, it is possible to construct a fractional CDS packing of size $\Omega(kp^2/\log^3 n)$. Since the size of any (fractional) CDS packing of a graph is upper bounded by its vertex connectivity, this directly implies that the vertex connectivity of the remaining graph is also at least $\Omega(kp^2/\log^3 n)$. While two of the logarithmic factors in this approach seem to be artifacts of the details in the method, the third one appears to be an inherent limitation of the method. This is because [6] shows that there are graphs with vertex connectivity k that have maximum (fractional) CDS packing size of $O(k/\log n)$. Thus, the approach of using a CDS packing as a witness for the vertex connectivity of the sampled subgraph inherently cannot prove a bound better than $\Omega(kp^2/\log n)$.

The New Approach. Our main contribution is a new analysis that establishes a tight lower bound on the vertex-sampling probability p that preserves simple connectivity. Moreover, we provide a new method that obtains a lower bound on the remaining vertex connectivity after node sampling, which relies on the bound for simple connectivity. A similar method yields a lower bound on the remaining vertex connectivity after edge sampling.

The formal statement of the lower bound on the vertex-sampling probability that preserves simple connectivity is as follows.

Theorem 3. *Let $G = (V, E)$ be a k-vertex-connected n-node graph. For an arbitrary $0 < \delta < 1$, let S be a randomly sampled subset of V such that each $v \in V$ is included in S independently with probability $p \geq \beta\sqrt{\log(n/\delta)/k}$, for a sufficiently large constant β (independent of n, k and δ). Then, with probability at least $1 - \delta$, the set S is a connected dominating set of G (and thus graph $G[S]$ is connected).*

The sampling probability threshold in the above theorem is expressed in terms of the probability $1 - \delta$ by which we require that the sampled set be connected. This is an important feature of the theorem, as we will need to plug exponentially small error probabilities δ to derive our main results.

A key idea in the proof of Theorem 3 is the novel notion of *λ-semi-connectivity*, which allows for a more refined analysis than by working directly with connectivity. We call a vertex set $S \subseteq V$ λ-semi-connected if for every partition of the connected components of $G[S]$ into two parts, there are λ nodes in $V \setminus S$ which are adjacent to components on both sides of the partition. We observe that if we start with a λ-semi-connected set S of already sampled nodes, where $\lambda = \Theta(\sqrt{k\log(n/\delta)})$, then sampling the remaining nodes with probability $p = \Theta(\log(n/\delta)/\lambda) = \Theta(\sqrt{\log(n/\delta)/k})$ suffices to obtain a connected dominating set with the required probability $1 - \delta$. This can be easily shown using techniques similar to those in [1, 6]. The difficult part is to prove that the above sampling probability p suffices to achieve the desired λ-semi-connectivity in the first place. This is the main technical contribution of our paper, and is shown by carefully analyzing how the semi-connectivity of the sampled set grows by adding new random vertices. In particular, we describe an *edge-coloring* procedure that takes place along with the sampling process, and we look at a new notion

[2] A dominating set is a set of nodes such that each node not in this set is adjacent to some node from the set. A graph with vertex connectivity k has minimum degree at least k, and thus a dominating set is already obtained w.h.p. when sampling with probability $\Omega(\log(n)/k)$.

[3] A *fractional CDS packing* is a collection of CDSs, each having a weight in the range $(0, 1]$, such that for every node, the sum of weights of all CDSs to which it belongs does not exceed 1. The size of a packing is the total weight of all CDSs in the collection. The case of all weights being 1 is simply a *CDS packing* and its size is the number of CDSs.

of connectivity between sampled nodes, which we call *i-novo-connectivity* (for colors $i = 1, \ldots, \lambda$). The connection between semi-connectivity and novo-connectivity is that once all sampled nodes are part of a single i-novo-connected component, the set of sampled nodes is i-semi-connected. In particular, we bound the number of sampling rounds required to obtain a single λ-novo-connected component, which then gives the desired property of λ-semi-connectivity.

Theorems 1 and 2 follow relatively easily, once we have established Theorem 3. To show the lower bound of Theorem 1, on the remaining vertex connectivity after vertex sampling, we view the sampling with probability p as a two-step process: first sampling with probability $2p$, and then subsampling with probability $1/2$. We argue that if the set sampled in the first step did not have sufficiently high vertex connectivity (with sufficiently high probability), then the two-step sampling would not result in a connected set with sufficiently high probability, thus contradicting Theorem 3.

The proof of the lower bound of Theorem 2, on the remaining vertex connectivity after edge sampling, is along similar lines.[4] We consider a two-phase sampling process: in one phase *edges* are sampled with probability p, and in the other phase *vertices* are sampled with probability $1/2$. We can analyze the process in two ways. One way is first to argue that vertex sampling with probability $1/2$ reduces vertex connectivity by at most a constant factor, by Theorem 1, implying the same lower bound on edge connectivity, and then to apply an edge sampling result by Karger [10] to conclude that connectivity is preserved with very high probability. The second way of analyzing the process is first to bound from below the remaining vertex connectivity after edge sampling, as in the statement of Theorem 2, and to combine that with the probability that a minimum vertex cut in the sampled graph survives the subsequent vertex sampling. Comparing the results of the above two approaches yields the bound of Theorem 2.

1.2 Optimality of our Results

The bound of Theorem 1 is existentially tight up to constant factors, as demonstrated in the following simple example.[5]

Proposition 4. *Let G be a $2n$-node graph consisting of two disjoint n-node cliques connected via a matching of $k \leq n$ edges. The vertex connectivity of G is k, and when each node is sampled with probability $p \geq 2 \ln n/n$, the expected vertex connectivity of the subgraph induced by the sampled nodes is at most $kp^2 + o(kp^2)$. If the sampling probability is $p = o(\sqrt{\log(n)/k})$, then the subgraph is disconnected[6] with probability at least $n^{-o(1)}$.*

Even if one desires the sampled subgraph to be connected with merely a *constant* probability, our vertex sampling threshold $p = \Omega(\sqrt{\log(n)/k})$ is essentially tight as shown by the next simple example.

Proposition 5. *Let G be an n-node graph consisting of n/k k-cliques ordered 1 to n/k, where each two consecutive cliques are connected via a k-edge matching. We assume that n is a multiple of k, and $k < n$. Graph G has vertex connectivity k, and when sampling nodes with probability $\omega(1/n) < p < o(\sqrt{\log(n/k)/k})$, the subgraph induced by the sampled nodes is disconnected with probability $1 - o(1)$.*

The bound of Theorem 2 is also existentially optimal: Karger's result from [10] states that the remaining edge connectivity is $\Theta(\lambda p)$ w.h.p. after edge sampling with probability p, when the initial edge connectivity is λ. Since the vertex connectivity k of a graph is upper bounded by its edge connectivity λ, but there are also graphs with $k = \lambda$, Karger's result implies, for such graphs, that the remaining vertex connectivity is at most at most $O(\lambda p)$ w.h.p. after edge sampling with probability p.

[4] We thank an anonymous reviewer of this journal paper for suggesting to us this simpler proof instead of the one we had initially.

[5] The proofs of Propositions 4 and 5 are given in Section 4.

[6] For the purposes of this statement, we consider the empty graph disconnected.

1.3 Implications

The fact that Theorem 1 proves an $\Omega(k)$ remaining vertex connectivity when $p = 1/2$, combined with the approach in [6, Section 5], imply the following corollary.

Corollary 6. *Any k-vertex-connected n-node graph can be decomposed into $\Omega(k/\log^2 n)$ vertex-disjoint connected dominating sets (CDS).*

This improves over the $\Omega(k/\log^5 n)$ bound of [6, Theorem 1.2]. As explained in [6, 5], decomposing to vertex-disjoint connected dominating sets can be viewed as a decomposition of vertex connectivity. This makes Corollary 6 the best known counterpart of the famous results of Tutte [16] and Nash-Williams [14] from 1961 for decomposing edge connectivity; namely that each λ-edge-connected graph contains $\lceil \frac{\lambda-1}{2} \rceil$ edge-disjoint spanning trees. The $\Omega(k/\log^2 n)$ bound of Corollary 6 is within an $O(\log n)$ factor of optimal because, as shown in [6], there exist k-connected graphs that cannot be decomposed into more than $\Theta(k/\log n)$ vertex disjoint connected dominating sets. Furthermore, the decomposition stated in Corollary 6 can be computed very efficiently, namely in $\tilde{O}(m)$ time where m is the number of edges in the graph, by combining random sampling with the approach of [5].

In addition, following the connection stated in [5, Section 1.4.1], Corollary 6 implies the best known approximation of the 1989 conjecture of Zehavi and Itai [17]. This conjecture states that each k-vertex-connected graph contains k *vertex-independent trees*, that is, k spanning trees rooted in a node $r \in V$ such that for each vertex $v \in V$, the paths between r and v in different trees are internally vertex-disjoint. We get the following approximation.

Corollary 7. *Any k-vertex-connected n-node graph contains $\Omega(k/\log^2 n)$ vertex-independent trees.*

2 Proofs of Theorems 1 and 2 Assuming Theorem 3

In this section, we prove Theorem 1 and Theorem 2, assuming that Theorem 3 holds. In simple words, the arguments presented in this section allow us to turn a "*very high probability of remaining (simply) connected after sampling*" to a "*very likely high vertex connectivity after sampling*."

2.1 Proof of Theorem 1: Vertex Connectivity under Vertex Sampling

Let $\alpha = 2\beta\sqrt{3}$, where β is the constant in the statement of Theorem 3, and let $\gamma = \alpha^{-2}$. We will show that for vertex-sampling probability $p \geq \alpha\sqrt{\log(n)/k}$, the sampled set S is a dominating set and $G[S]$ has vertex connectivity at least $\gamma k p^2$, with probability at least $1 - 2^{-\gamma k p^2}$.

Assume, towards a contradiction, that the above is not true. That is, for some sampling probability $p \geq \alpha\sqrt{\log(n)/k}$, with probability greater than $2^{-\gamma k p^2}$, the set S is not a dominating set or $G[S]$ has vertex connectivity less than $\gamma k p^2$. We show that this contradicts Theorem 3.

Consider a vertex sampling of G with sampling probability $q = p/2$, executed in two steps: the first step is sampling with probability p, and the second is an independent subsampling with probability $1/2$.

Let $\mathcal{E}$ denote the event that after sampling with probability p, the sampled set S is not a dominating set or $G[S]$ has vertex connectivity less than $\gamma k p^2$. Then $\Pr(\mathcal{E}) > 2^{-\gamma k p^2}$, by our assumption above. Suppose event $\mathcal{E}$ holds and S is a dominating set, thus $G[S]$ must have vertex connectivity less than $\gamma k p^2$, and consider a single vertex cut in $G[S]$ of size at most $\gamma k p^2$. During the further $1/2$-subsampling, all nodes of this cut are removed with probability at least $2^{-\gamma k p^2}$, and if this happens then the final sampled set is not a connected dominating set. Combining this with $\Pr(\mathcal{E}) > 2^{-\gamma k p^2}$, we obtain that vertex sampling of G with probability q has probability more than $2^{-2\gamma k p^2}$ to sample a set that is not a connected dominating set.

On the other hand, by applying Theorem 3 with failure probability $\delta = 2^{-2\gamma kp^2}$, we obtain that vertex sampling of G with probability at least $\beta\sqrt{\log(n/\delta)/k}$, yields a connected dominating set with probability at least $1-\delta = 1-2^{-2\gamma kp^2}$. However, since

$$\beta\sqrt{\frac{\log(n/\delta)}{k}} = \beta\sqrt{\frac{\log n}{k}+2\gamma p^2} = \beta\sqrt{\gamma}p\sqrt{\frac{\log n}{k\gamma p^2}+2} = \beta\alpha^{-1}p\sqrt{\frac{\log n}{k\alpha^{-2}p^2}+2} \leq \beta\alpha^{-1}p\sqrt{3} = \frac{p}{2} = q,$$

this contradicts the result we showed just before.

2.2 Proof of Theorem 2: Vertex Connectivity under Edge Sampling

We will use the following two results. The first is by Karger [10], and the second is a corollary of Theorem 1.

Proposition 8 ([10])**.** *There are constants $\zeta, \eta > 0$, such that for any n-node λ-edge-connected graph, independent edge sampling with probability $p \geq \zeta\log(n)/\lambda$ yields a spanning subgraph with edge connectivity at least $\eta\lambda p$, with probability at least $1-2^{-\eta\lambda p}$.*

Proposition 9. *There are constants $g, h > 0$, such that for any n-node graph with vertex connectivity $k \geq g\log n$, vertex sampling with probability $1/2$ yields a dominating set which induces a subgraph with edge connectivity at least hk, with probability at least $1-2^{-hk}$.*

Proposition 9 follows from Theorem 1, by letting $p = 1/2$, and using the fact that the edge connectivity of a graph is greater than or equal to its vertex connectivity.

Let $\alpha = \zeta/h$ and $\gamma = \eta h/2$, and assume w.l.o.g. that $\eta \leq 1$. Suppose also that $k \geq g\log n$ (otherwise Theorem 2 holds trivially). We will show that for edge-sampling probability $p \geq \alpha\log(n)/k$, the sampled subgraph G' has vertex connectivity at least γkp, with probability at least $1-2^{-\gamma kp+1}$.

Assume for contradiction that this is not true. Then for some edge-sampling probability $p \geq \alpha\log(n)/k$, with probability greater than $2^{-\gamma kp+1}$ the sampled graph G' has vertex connectivity less than γkp.

Consider the following two-phase sampling process on G: First we sample *edges* with probability p, and then in the resulting subgraph G' we sample *vertices* with probability $1/2$. Let $\mathcal{E}$ be the event that the vertex connectivity of G' is less than γkp; then $\Pr(\mathcal{E}) > 2^{-\gamma kp+1}$, by our assumption above. Suppose that event $\mathcal{E}$ holds, and consider a single vertex cut in G' of size at most γkp. Then in the vertex-sampling phase with probability $1/2$, all nodes of this cut are removed with probability at least $2^{-\gamma kp}$, and if this happens then the final sampled set of vertices is not a connected dominating set. Combining this with $\Pr(\mathcal{E}) > 2^{-\gamma kp+1}$, we obtain that the two-phase sampling process has probability more than $2^{-2\gamma kp+1}$ to sample a vertex set that is not a connected dominating set of G.

Consider now the same sampling process, but with the two phases executed in reverse order: First we sample vertices with probability $1/2$, and if S is the sampled set, we then sample edges from $G[S]$ with probability p, obtaining a subgraph H of $G[S]$. The outcome of this sampling process is the vertex set of H. From Proposition 9, it follows that S is a dominating set and $G[S]$ has edge connectivity at least hk, with probability at least $1-2^{-hk}$. And from Proposition 8, it follows that, if $G[S]$ has edge connectivity at least $\lambda = hk$, then H is a spanning subgraph of $G[S]$, with probability at least $1-2^{-\eta hkp}$ (here we used that $p \geq \alpha\log(n)/k = \zeta\log(n)/\lambda$, which allows us to apply Proposition 8). Combining these two results, we obtain that the two-phase sampling process has probability at least $1-2^{-hk}-2^{-\eta hkp}$ to result in a subgraph H with vertex set S which is a connected dominating set of G.

Observe now that the outcome of the two-phase sampling process should not depend on the order in which the phases are executed. Comparing then the results above for the two different orders of the phases, we reach the desired contradiction, because $2^{-hk}+2^{-\eta hkp} \leq 2^{-2\gamma kp+1}$. This inequality is obtained by using that $\gamma = \eta h/2$ and $\eta \leq 1$.

3 Proof of Theorem 3: Simple Connectivity under Vertex Sampling

In this section we prove our main technical result, Theorem 3, which establishes a lower bound on the vertex-sampling probability that preserves (simple) connectivity with a given probability $1 - \delta$.

In Section 3.1, we formally define the notion of semi-connectivity and prove Theorem 3 using a key lemma on the sampling probability needed to achieve a certain degree of semi-connectivity. This lemma is at the core of our analysis and is proven in Section 3.2, by introducing the notion of novo-connectivity and a related edge-coloring process.

3.1 Proof of Theorem 3 via Semi-Connectivity

We start by fixing some basic notation, and defining the key notion of semi-connectivity. We say that a node $u \in V$ is a *neighbor of* $S \subseteq V$ or is *adjacent to* S if u is adjacent to some node $v \in S$ and $u \notin S$. The set of neighbors of S is denoted by ∂S. An edge or path *between two sets* S *and* S' is one with endpoints $u \in S$ and $u' \in S'$.

Definition 10 (λ-Semi-Connected Set)**.** A vertex set $S \subseteq V$ is λ-*semi-connected*, for some $\lambda \geq 0$, if for any partition of S into two sets T and $S \setminus T$ with no edges between them, T and $S \setminus T$ have at least λ common neighbors, i.e., $|\partial T \cap \partial(S \setminus T)| \geq \lambda$.

If a set is $(\lambda + 1)$-semi-connected, then it is λ-semi-connected, as well. Also, any connected set is λ-semi-connected for any $\lambda \geq 0$, as the condition in Definition 10 is vacuously true in this case.

Before we proceed to the detailed analysis we provide first some intuition on how semi-connectivity is relevant to the problem we try to solve.

Intuition for Semi-Connectivity. Consider a natural interpretation of sampling that was introduced for the problem in [6], in which one looks at the sampling process as slowly adding nodes over time. In particular, instead of sampling nodes with probability p at once, one samples nodes over multiple, $T = \Omega(\log n)$, rounds, where in each round nodes are sampled with some smaller probability $q \approx p/T$. This allows to study and analyze the emergence and merging of connected components, as time progresses and more and more nodes are sampled.

Let us take a look at a single edge-cut, the canonical bad cut consisting of a k-edge-matching as discussed in Proposition 4. We emphasize that understanding the behavior of *all* cuts simultaneously is the part that makes the problem challenging, but focusing on this single cut should be sufficient for delivering the right intuition about the key new element in our analysis.

In the cut consisting of a k-edge-matching, in any round, both endpoints of an edge will become sampled with probability q^2. Since there are k such edges, the probability that at least one edge gets sampled in a round is bounded by kq^2. Now, in order for at least one edge of the cut to be sampled w.h.p. in this way over the course of T rounds, we need that $Tkq^2 = kp^2/T = \Omega(\log n)$. Since we assumed $T = \Omega(\log n)$, this results in $p = \Omega(\log(n)/\sqrt{k})$ being a necessary condition. This explains in a very simplified manner why the argument in [6] does not work for $p = o(\log(n)/\sqrt{k})$.

Here, we refine this layer-by-layer sampling by further exploiting that connectivity evolves gradually. In particular, while the probability of obtaining one complete edge in one round is only q^2, and thus quite small, the number of sampled nodes on each side of the cut grows by roughly kq in each round. Thus, after λ/kq rounds for some $\lambda = \Omega(\log n)$, the number of such nodes is at least λ w.h.p. Each of these nodes intuitively already goes half way in crossing the cut. In particular, with λ such nodes, there is a chance of λq per each of the next rounds to complete such a semi-sampled edge into a fully sampled edge that crosses the cut. This means that after such λ-"semi-connectivity" is achieved, w.h.p. no more than $\log(n)/\lambda q$ further rounds are needed to get an edge crossing the cut to be fully sampled. The optimal value for λ is now chosen

to balance between the λ/kq rounds to achieve λ-semi-connectivity and the $\log(n)/\lambda q$ additional rounds required to achieve connectivity. This leads to $\lambda = \sqrt{\log(n)/k}$, and results in $T = \Theta\big(\sqrt{\log(n)/k}/q\big)$ rounds and a sampling probability of $p = \Theta\big(\sqrt{\log(n)/k}\big)$ being sufficient for a single cut.

In the above description we focused on a single cut. Understanding, however, the behavior of all (the exponentially many) cuts together turns out to be significantly more complex. Overall, the main technical challenge in this paper is to develop notions, definitions, and arguments to prove that semi-connectivity indeed gets established quickly, for all cuts.

Detailed Analysis. We now describe in detail how to obtain Theorem 3 by analyzing semi-connectivity. At a high level, the process of sampling consists of three parts for obtaining: (i) domination, (ii) λ-semi-connectivity for a $\lambda = \Theta(\sqrt{k \log(n/\delta)})$, and (iii) connectivity. Establishing domination is trivial, and the proof of connectivity after having λ-semi-connectivity follows easily from the layer-by-layer analysis of [6]. The key challenge is to prove λ-semi-connectivity. Precisely, we show that sampling with probability $\Theta(\lambda/k)$ suffices to increase the semi-connectivity of a dominating set by an additive term of λ, for $\lambda = \Omega(\log n)$.

We start by the simple observation that adding a node from $V \setminus S$ to a λ-semi-connected set S does not break semi-connectivity, provided that S is a dominating set.

Claim 11. *If $S \subseteq V$ is a λ-semi-connected dominating set, then for any node $u \in V \setminus S$, the set $S \cup \{u\}$ is also a λ-semi-connected dominating set.*

Proof. Since S is a dominating set, so is the set $S' = S \cup \{u\}$. Next, we show that S' is λ-semi-connected. Consider any partition of S' into two sets T' and $S' \setminus T'$, such that these sets have no edges between them. We show that T' and $S' \setminus T'$ have at least λ common neighbors. We assume w.l.o.g. that $u \in T'$. We observe that $T' \neq \{u\}$, because S is a dominating set and thus if $T' = \{u\}$ then there would be an edge between T' and $S' \setminus T'$. Thus, the set $T = T' \setminus \{u\}$ is non-empty, and the two sets T and $S \setminus T = S' \setminus T'$ constitute a partition of S. We have that T and $S \setminus T$ have no edges between them, for otherwise the same edge would also connect T' and $S' \setminus T'$. Since S is λ-semi-connected, there are at least λ common neighbors for T and $S \setminus T$. Each of these nodes is also a common neighbor for T' and $S' \setminus T'$, because it cannot be equal to u (otherwise there is an edge between T' and $S' \setminus T'$). Since this holds for any such partition, this implies that S' is λ-semi-connected. □

Next we show that, if we start with a set S of nodes that is a λ-semi-connected dominating set, then it suffices to sample the remaining nodes with probability $\Theta(\log(n/\delta)/\lambda)$ to end up with a connected dominating set with probability $1 - \delta$.

Lemma 12. *Let $S \subseteq V$ be a λ-semi-connected dominating set. Sampling each remaining node $u \in V \setminus S$ with probability $\log_\gamma(n/\delta)/\lambda$, where $\gamma = \frac{2e}{e+1}$, yields a set S' such that $S \cup S'$ is a connected dominating set with probability at least $1 - \delta$.*

Proof. We perform sampling in rounds, where in each round every node that has not been sampled yet is sampled with probability $1/\lambda$. The total number of rounds is $r = \log_\gamma(n/\delta)$, thus the probability for any given node $u \in V \setminus S$ to be sampled in one of those rounds is at most $r/\lambda = \log_\gamma(n/\delta)/\lambda$, as required by the lemma statement. Let S_i, for $0 \leq i \leq r$, denote the set consisting of all nodes sampled in the first i rounds and all $u \in S$ (so $S_0 = S$). Further, let X_i denote the number of connected components of the induced subgraph $G[S_i]$. We bound $\mathbf{E}[X_i]$ next.

Fix set S_i and suppose that $G[S_i]$ is disconnected, i.e., $X_i > 1$. Since S is a λ-semi-connected dominating set, S_i is also a λ-semi-connected dominating set, by Claim 11. Hence, each connected component C of $G[S_i]$ has at least λ common neighbors with other connected components. If any of those common

neighbors gets sampled in round $i+1$, then C is merged with another component. Then the probability of C to get merged in round $i+1$ is at least $1-(1-1/\lambda)^\lambda \geq 1-1/e$. Since the drop $X_i - X_{i+1}$ in the number of connected components in round $i+1$ is at least half the total number of connected components that get merged with another component, it follows that

$$\mathbf{E}[X_i - X_{i+1} \mid S_i] \geq \frac{1-1/e}{2} \cdot X_i = (1-1/\gamma)X_i.$$

This inequality assumes that $X_i > 1$ (notice that X_i is fixed because S_i is fixed). To lift this assumption we define the random variables $Y_i = X_i - 1$ and work with them instead. We have

$$\mathbf{E}[Y_i - Y_{i+1} \mid S_i] = \mathbf{E}[X_i - X_{i+1} \mid S_i] \geq (1-1/\gamma)X_i \geq (1-1/\gamma)Y_i.$$

The above inequality $\mathbf{E}[Y_i - Y_{i+1} \mid S_i] \geq (1-1/\gamma)Y_i$ also holds (trivially) when $X_i = 1$, since then $Y_i = 0$. Taking now the unconditional expectation yields $\mathbf{E}[Y_i - Y_{i+1}] \geq (1-1/\gamma)\,\mathbf{E}[Y_i]$, which implies $\mathbf{E}[Y_{i+1}] \leq \mathbf{E}[Y_i]/\gamma$. Applying this inequality repeatedly gives

$$\mathbf{E}[Y_r] \leq \mathbf{E}[Y_0]/\gamma^r \leq n/\gamma^r,$$

since $Y_0 < n$. This yields $\mathbf{E}[Y_r] \leq n/\gamma^r = \delta$, as $r = \log_\gamma(n/\delta)$. By Markov's inequality then we obtain $\Pr(Y_r > 0) = \Pr(Y_r \geq 1) \leq \mathbf{E}[Y_r]/1 \leq \delta$. Therefore, the probability that there is only one connected component by the end of the last round is at least $1-\delta$. □

Lemma 12 requires that we start with a set S of (already sampled) nodes which is a λ-semi-connected dominating set. To achieve domination (but not λ-semi-connectivity) with probability at least $1-\delta$, it suffices to sample nodes with probability $\Theta(\log(n/\delta)/k)$ (recall, k is the vertex connectivity of the graph):

Lemma 13. *Sampling each node with probability $\ln(n/\delta)/k$ yields a dominating set with probability at least $1-\delta$.*

Proof. From the k-vertex-connectivity of the graph, it follows that each node has degree at least k. Thus the probability for a given node that none of its neighbors gets sampled is at most $\left(1-\frac{\ln(n/\delta)}{k}\right)^k \leq e^{-\frac{\ln(n/\delta)}{k}\cdot k} = \frac{\delta}{n}$. By the union bound, the probability that this happens for at least one of the n nodes is at most δ. □

It remains to bound the sampling probability needed to achieve λ-semi-connectivity. This is the key part in our analysis. In particular, we show that a sampling probability of $\Theta((\lambda+\log n)/k)$ suffices to achieve λ-semi-connectivity. Section 3.2 is dedicated to the proof of this result, which is formally stated as follows.

Lemma 14 (Key Semi-Connectivity Claim)**.** *Let $S \subseteq V$ be a dominating set. Sampling each remaining node $u \in V \setminus S$ with probability $16\lambda/k$ yields a set S' such that $S \cup S'$ is a λ-semi-connected dominating set with probability at least $1 - n/2^\lambda$.*

We now have all the ingredients to prove Theorem 3.

Proof of Theorem 3. If $k = O(\log(n/\delta))$ then the theorem holds trivially by choosing the constant β such that $\beta\sqrt{\log(n/\delta)/k} \geq 1$. Below we assume that $k > \log(3n/\delta)$.

We consider three phases. First, we sample nodes with probability $\ln(3n/\delta)/k$, and from Lemma 13 we have that the resulting set, denoted S_1, is a dominating set with probability $1-\delta/3$.

In the next phase, we sample the remaining nodes $u \in V \setminus S_1$ with probability $16\lambda/k$, for $\lambda = \sqrt{k\log(3n/\delta)}$. From Lemma 14 it follows that if S_1 is a dominating set, then the set S_2 of all nodes sampled in the first two phases is a λ-semi-connected dominating set with probability $1 - n/2^\lambda$. Note that $1 - n/2^\lambda \geq 1 - \delta/3$, because $\lambda = \sqrt{k\log(3n/\delta)} \geq \log(3n/\delta)$, as we have assumed $k > \log(3n/\delta)$.

In the last phase, we sample the remaining nodes $u \in V \setminus S_2$ with probability $\log_\gamma(n/\delta)/\lambda$, and obtain from Lemma 12 that the probability for the set S_3 of nodes sampled in the three phases to be a connected dominating set is at least $1 - \delta/3$, provided that S_2 is a λ-semi-connected dominating set.

A union bound over all three phases shows that the probability of ending up with a connected dominating set S_3 is indeed $1 - \delta$, and the total sampling probability is at most

$$\frac{\ln(3n/\delta)}{k} + \frac{16\sqrt{k\log(3n/\delta)}}{k} + \frac{\log_\gamma(n/\delta)}{\sqrt{k\log(3n/\delta)}},$$

which is $O\big(\sqrt{\log(n/\delta)/k}\big)$. □

3.2 Proof of Lemma 14: Sampling Threshold for λ-Semi-Connectivity

We assume that sampling is performed in rounds. In each round, each node not sampled yet is sampled with probability $1/k$. Within a round, the sampling of nodes is done sequentially, in *steps*, with a single node considered for sampling at each step (the order in which nodes are considered in a round can be arbitrary). We will denote by S_t the set containing all nodes sampled in the first t steps and all nodes $u \in S$ (so $S_0 = S$). To simplify notation we will say that the nodes $u \in S$ were also 'sampled', before the first step.

Along with the sampling process, we consider a procedure that colors the *edges* of the graph. We describe this procedure and the related notion of *novo-connectivity* next.

Edge-Coloring Procedure. At any point in time, each edge has a color from the set {black, gray, white, color-1, ..., color-λ}. The same color can be used for more than one edge, and the color of an edge may change during the sampling process.

We have the following coloring initially: Edges with both endpoints in $S_0 = S$ are black; the edges between S and $V \setminus S$ are gray; and all remaining edges (between nodes from $V \setminus S$) are white. There are no color-i edges initially, for $1 \le i \le \lambda$.

In each step of the sampling process, some edges may change color. The possible changes are that white edges may switch to color-i, for some i, and edges of any color may switch to black. At any point in time we have the following invariants:

- An edge is black iff both its endpoints belong to the set S_t of nodes sampled up to that point.
- If an edge is gray or of color-i, for some i, then exactly one of its endpoints is in S_t and the other in $V \setminus S_t$.
- If both endpoints of an edge are in $V \setminus S_t$ then this edge is white. (But it is possible for a white edge to have one endpoint in S_t and the other in $V \setminus S_t$.)

Before we describe precisely the color changes that take place in each step we must introduce the key concept of *novo-connectivity*. In the following definition, a path is not necessary simple, i.e., it may visit the same vertex more than once.

Definition 15 (i-Novo-Connectivity)**.** A path between two *sampled* nodes is an *i-novo-path*, for some $1 \le i \le \lambda$, if (1) each edge along the path has a color from the set {black, gray, color-i}, and (2) for any two consecutive edges whose common endpoint is not sampled, at least one of them is a color-i edge (the other edge is then either color-i or gray). Two sampled vertices are *i-novo-connected* if there is a (not necessarily simple) i-novo-path between them. An *i-novo-connected component*, or simply *i-novo-component*, is a maximal subset of the sampled nodes such that any two nodes in that set are i-novo-connected.

The definition of i-novo-connectivity does not require that two i-novo-connected nodes have a *simple* i-novo-path between them. Moreover, this is not implied by Definition 15: e.g., consider a non-simple i-novo-path $uxyzxv$, where all nodes except for x are sampled, edges ux and xv are gray, and edges xy and zx are color-i; the simple path uxv is not an i-novo-path, because it does not satisfy Condition (2). Nevertheless, for the specific rules we use for updating the edge colors, described below, it can be shown that a simple i-novo-path exists between any two i-novo-connected nodes. This result is not needed for our analysis, but is an interesting property, which may be useful for other application of this technique. For that reason, the proof of this result is given in the appendix (see Lemma 28).

We now describe the color changes that take place during step $t \geq 1$. Suppose that node $u \notin S_{t-1}$ is considered for sampling in step t. If u is not sampled in that step, i.e., $S_t = S_{t-1}$, then there are no color changes. If u is sampled, i.e., $S_t = S_{t-1} \cup \{u\}$, all edges uv with $v \in S_{t-1}$ become black, and then the following λ *sub-steps* are performed. In each sub-step $i = 1, \ldots, \lambda$, some edges incident to u may switch from white to color-i. Precisely, an edge uv switches to color-i in sub-step i of step t if all the conditions below hold simultaneously:

1. uv is white before sub-step i.
2. v is adjacent to only one i-novo-component at the beginning of step t—we say v is an *exclusive* neighbor of that component.
3. u is not adjacent to the same i-novo-component as v at the beginning of step t.

We also have the additional rule:

4. If there are more than one node v that satisfy the three conditions above and are adjacent to the *same* i-novo-component at the beginning of step t, then only one edge uv is colored with color-i (choosing an arbitrary one among those nodes v).

Intuition for Novo-Connectivity. We provide now some intuition on how the notion of novo-connectivity and the edge-coloring procedure defined above are used to establish the sampling probability threshold for λ-semi-connectivity described in Lemma 14.

We consider how i-novo-components evolve over time as sampling proceeds. First, we argue that eventually all sampled nodes belong to a single i-novo-component, for any $i \in \{1, \ldots, \lambda\}$. Then we show how this implies that the set of sampled nodes is λ-semi-connected.

Initially, when there are only black, white, and grey edges, the i-novo-components are precisely the connected components of $G[S]$. As more nodes are sampled, the i-novo-components expand and also merge with other i-novo-components.

To obtain a lower bound on the rate at which the number of i-novo components drops, we focus on just two ways in which two i-novo-components can merge into a single component: (1) a common neighbor of them gets sampled; (2) a neighbor u of the one component gets sampled, and then an edge between u and an exclusive neighbor v of the second component gets colored color-i. In the latter, the requirement of the edge-coloring procedure that v must be an exclusive neighbor of the second component in order for uv to get colored implies (together with the fact that S is a dominating set) that v has a gray edge to some node w of that component. Therefore, an i-novo-path uvw is created between u and the second i-novo-component.

Based on the above two ways of merging, we compute a lower bound on the probability that a given i-novo-component C merges with another i-novo-component in a round. For that we identify a set of 'critical' nodes for C, such that sampling any of those nodes would result in C being merged with another i-novo-component. From the k-vertex-connectivity of G, it follows that there are at least k internally-disjoint paths from C to other i-novo-components. By taking the k shortest such internally-disjoint paths, we can easily

argue (using that S is a dominating set) that each of these paths has length 2 or 3: For a path of length 2, the internal node is a common neighbor of C and another i-novo-component; for a path of length 3, the first internal node is an exclusive neighbor of C, and the second a neighbor of another i-novo-component. Call the single internal node of each of the paths above with length 2 a 1-*critical* node for C, and the second internal node of each path with length 3 a 2-*critical* node for C; recall that all these k nodes are disjoint.

Since each of the critical nodes for C gets sampled independently with probability $1/k$ in a round, it follows that at least one of them is sampled with probability $1-(1-1/k)^k \geq 1-1/e$. We would like to argue that if any of these nodes gets sampled then C gets merged. This is definitely the case when a 1-critical node gets sampled. However, this is not always the case when some 2-critical node u for C gets sampled. The reason is that in the latter case, u may also be a 2-critical node for a j-novo-component C', for some $j \neq i$, such that for both C and C', we need to color the same edge uv with color-i or color-j, respectively, in order for the corresponding component to get merged. However, the edge can get only one of the two colors, and the rules of the edge-coloring procedure give precedence to color-j over color-i in a round, if $j < i$ (as sub-step j is executed before sub-step i).

This has two implications: First, by giving precedence to smaller colors, we obtain that any two i-novo-connected nodes are also j-novo-connected, if $j < i$. (We use this later to argue that if all sampled nodes belong to a single λ-novo-component, they also belong to a single i-novo-component, for any $i \leq \lambda$; we also use it again at the end of the proof.) The second implication we establish is that if a 2-critical node u for C gets sampled but C does not get merged, then some distinct j-novo-component gets merged instead, for some $j < i$.

If it were the case that sampling a critical node for an i-novo-component C would always result in C being merged with another i-novo-component, we could easily conclude that after $\Theta(\log n)$ rounds (and a total sampling probability of $\Theta(\log(n)/k)$), there would be just a single i-novo-component: in each round, the number of i-novo-components drops by a constant factor in expectation. But instead, we have that sampling a critical node for an i-novo-component C only ensures that some distinct j-novo-component, where $j \leq i$, gets merged. We provide a more refined argument establishing that in a round, the expected decrease in the total number of j-novo-components with $j \leq i$, is bounded from below by a linear function of the current number of i-novo-components. This implies that after $\Theta(\log n + i)$ rounds there is just a single i-novo-component. Hence, for $\lambda = \Omega(\log n)$, there is just a single λ-novo-component after $\Theta(\lambda)$ rounds, and thus a single i-novo-component, for any $i \leq \lambda$.

We have now finished the informal presentation of the argument that eventually (i.e., after $\Theta(\lambda)$ rounds) all sampled nodes belong to a single i-novo-component, for any $i \leq \lambda$. We still need to explain how this yields the desired λ-semi-connectivity for the set of sampled nodes.

Suppose all sampled nodes belong to a single i-novo-component, for each $i \leq \lambda$. This implies that for any partition of the sampled nodes to sets A and B with no edges between them, there is at least one i-novo-path between A and B. From the assumption that S is a dominating set, it follows easily (using just the definition of i-novo-connectivity) that there is an i-novo-path between A and B that has length 2: one endpoint of this path is in A, the other in B, and its internal node is not sampled; one of the two edge of the path is color-i while the other is color-i or gray. To establish the desired λ-semi-connectivity of the set of sampled nodes, we argue that there are λ such paths of length 2, one for each $i \in \{1, \ldots, \lambda\}$, which are internally disjoint. This implies that A and B have (at least) λ common neighbors. In particular, we consider for each i, the *first* such path created. We sketch the argument next.

Let $u_i w_i v_i$ denote the first i novo-path of length 2 created between A and B (we assume $u_i \in A$ and $v_i \in B$). We must argue that all nodes w_i, $1 \leq i \leq \lambda$, are distinct. Suppose for contradiction that $w_i = w_j = w$, for some $j < i$. At the time when i-novo-path $u_i w v_i$ is created, one of its edges gets colored with color-i; suppose $w v_i$ is that edge. First observe that edge $w v_j$ must be color-j eventually: if not, it must be gray, as $u_j w v_j$ is an j-novo-path; but at the point when $w v_i$ gets colored color-i, w must be an exclusive neighbor of an i-novo-component, and if $w v_j$ is gray, it follows that u_i and v_j are

at the same i-novo-component at that time, and thus there was already an i-novo-path between $u_i \in A$ and $v_j \in B$, contradicting the assumption that $u_i w_i v_i$ was the first such path. Moreover, for the same reason v_i must have been sampled before v_j. This implies that i-novo-path $u_i w v_i$ was created before the j-novo-path $u_j w v_j$. Hence, the first i-novo-path between A and B was created before the first j-novo-path between them. However, this contradicts the earlier fact that if two nodes are i-novo-connected they are also j-novo-connected, for any $j < i$.

Road-map of the Rest of the Proof. The remainder of the proof of Lemma 14 unfolds in a series of claims. In Claim 16, we identify the set of i-novo-components that merge in a single step t, and then we prove that at any step each i-novo-component is a subset of an $(i-1)$-novo-component, in Claim 17. Next we introduce the notion of a critical node for an i-novo-component (Definition 18), and show that the number of critical nodes for each i-novo-component is at least equal to the vertex connectivity k, in Claim 19. In Claims 20–22 we show that the drop in the total number of j-novo-components in a round, for all $j \leq i$, is bounded from below by half the number of i-novo-components for which a critical node is sampled in the round. Then we bound from below the expected value of that drop using Claim 19, in Claims 23 and 24, and use this result in Claim 25 to bound by $O(\lambda)$ the number of rounds before there is just a single λ-novo-component. At that time, by Claim 17, there is just a single i-novo-component, for any $i \leq \lambda$. Finally, in Claim 27, we show that having just a single i-novo-component for each $i \leq \lambda$, implies λ-semi-connectivity, concluding the proof of Lemma 14.

Two distinct i-novo-components merge into a single i-novo-component if an i-novo-path is created between them. In a step, this happens when a common neighbor of the two components is sampled, or when a neighbor u of the one component is sampled and an edge uv to an exclusive neighbor v of the other component is colored with color-i, as explained in the next claim.

Claim 16. *Suppose that in step t node u is sampled, and $\mathcal{C}$ is the set of all i-novo-components C at the beginning of step t for which u is adjacent to C, or u is adjacent to an exclusive neighbor v of C and edge uv is colored with color-i in step t. Then all $C \in \mathcal{C}$ merge into a single i-novo-component $\left(\bigcup_{C \in \mathcal{C}} C\right) \cup \{u\}$ in step t, while the remaining i-novo-components do not change.*[7]

Proof. Let $u_1, \ldots, u_r \in S_{t-1}$ be the neighbors of u that are already sampled before step t. Let $D_1, \ldots, D_r$ denote the i-novo-components to which $u_1, \ldots, u_r$, respectively, belong to at the beginning of step t (these components are not necessarily distinct). When u is sampled, all edges uu_j, for $1 \leq j \leq r$, become black, and a new i-novo-component $D = \{u\} \cup D_1 \cup \cdots \cup D_r$ is formed, replacing $D_1, \ldots, D_r$. Other than that, no additional merges of i-novo-components occur before the first sub-step, since for any neighbor $v \notin S_{t-1}$ of u, edge uv was white before step t, and remains white until the first sub-step.

During the first $i-1$ sub-steps of step t, the i-novo-components do not change as the sets of black, gray, and color-i vertices do not change.

Consider now sub-step i, and suppose that edges $uv_1, \ldots, uv_\ell$ are colored with color-i in this sub-step. Let $1 \leq j \leq \ell$. From the edge-coloring procedure, it follows that v_j is an exclusive neighbor of some i-novo-component C_j at the beginning of step t (Rule 2), and u is not adjacent to C_j (Rule 3). Since u is not adjacent to C_j, the i-novo-component C_j does not change between the beginning of step t and the beginning of sub-step i, as we saw above.

We claim that an i-novo-path is created between u and C_j in sub-step i: Let w_j be a node in $C_j \cap S$ that is adjacent to v_j (recall that S is the set of nodes we start with, before the sampling). The node w_j exists, because S is a dominating set so there must be a node in S which is adjacent to v_j, and that node

[7] If $\mathcal{C}$ consists of a single i-novo-component C, then C is just replaced by $C \cup \{u\}$. It is not possible that $\mathcal{C} = \emptyset$, as S_{t-1} is a dominating set.

must belong to C_j because v_j is an exclusive neighbor of C_j. Therefore, the edge w_jv_j must be gray since $w_j \in S$ and $v_j \notin S_t$. Since the edge uv_j is colored with color-i, this implies an i-novo-path between u and w_j, and thus between u and C_j.

It follows that a new i-novo-component $D \cup C_1 \cup \cdots \cup C_\ell$ is formed in sub-step i, replacing $D, C_1, \ldots, C_\ell$. Other than that, no additional merges of i-novo-components occur in sub-step i: if an i-novo-component C is not adjacent to u or to some node v for which edge uv is colored in sub-step i, then no edges incident to C or to C's neighbors change color, thus no new i-novo-paths are created between C and other i-novo-components.

Finally, in the remaining sub-steps of step t after sub-step i the i-novo-components do not change.

The claim then follows by observing $\mathcal{C}$ consists precisely of the i-novo-components $D_1, \ldots, D_r$ and $C_1, \ldots, C_\ell$. □

Next we show that if two nodes are i-novo-connected, they are also $(i-1)$-novo-connected.

Claim 17. *At any step, each i-novo-component is a subset of some $(i-1)$-novo-component, for $2 \le i \le \lambda$.*

Proof. The proof is by induction on the number of steps t. The base case holds because when $t = 0$, there are only white, black, and gray edges, implying that an i-novo-component is also a j-novo-component, for any $i, j \in \{1, \ldots, \lambda\}$. Next we assume that the claim holds after the first $t-1$ steps and consider step t.

Suppose node u is sampled at step t, and let $\mathcal{C}_j$ be the set of all j-novo-components C at the beginning of step t, for which u is adjacent to C, or u is adjacent to an exclusive neighbor v of C and edge uv is colored with color-j in step t. From Claim 16, it follows that in step t, all i-novo-components $C \in \mathcal{C}_i$ merge into a single i-novo-component $A = \left(\bigcup_{C \in \mathcal{C}_i} C\right) \cup \{u\}$, and similarly all $(i-1)$-novo-components $C \in \mathcal{C}_{i-1}$ merge into a single $(i-1)$-novo-component $B = \left(\bigcup_{C \in \mathcal{C}_{i-1}} C\right) \cup \{u\}$, while the remaining i-novo-components and $(i-1)$-novo-components do not change in step t. Then to prove the claim it suffice to show that $A \subseteq B$. From the induction hypothesis, for each $C \in \mathcal{C}_i$, there is some $(i-1)$-novo-component $C' \supseteq C$ at the beginning of step t; we will show that $C' \in \mathcal{C}_{i-1}$. This then implies $A \subseteq B$.

If u is adjacent to C', then by definition $C' \in \mathcal{C}_{i-1}$, so suppose that u is not adjacent to C'. We must show that some edge uv' is colored with color-$(i-1)$, where v' is an exclusive neighbor of $(i-1)$-novo-component C' at the beginning of step t: Since u is not adjacent to C' and $C \subseteq C'$, u is not adjacent to C either. Then u is adjacent to some exclusive neighbor v of C, and edge uv is colored with color-i in sub-step i. We claim that edge uv fulfilled all the requirements for becoming color-$(i-1)$ in sub-step $i-1$ (Rules 1–3): uv was white before sub-step $i-1$, since it was white before sub-step i (otherwise uv would not be colored with color-i in sub-step i); v was an exclusive neighbor of $(i-1)$-novo-component C' at the beginning of step t, since it was an exclusive neighbor of i-novo-component C at the beginning of step t, and $C \subseteq C'$; and we have assumed that u is not adjacent to C'.

Since edge uv was not colored in sub-step $i-1$, despite satisfying the above requirements, it must be that by Rule 4, some other edge uv' was colored with color-$(i-1)$ in sub-step $i-1$, where v' was also an exclusive neighbor of $(i-1)$-novo-component C' at the beginning of step t. Therefore, $C' \in \mathcal{C}_{i-1}$. □

Next we define the notion of a *critical node* for an i-novo-component, and show that each i-novo-component has at least k such critical nodes, where k is the vertex connectivity of the graph.

Definition 18 (Critical Nodes)**.** Let C be an i-novo-component at the beginning of round r. A node u is *critical* for C in round r, if it is not sampled before round r, and one of the following two conditions holds at the beginning of round r: (1) u is a non-exclusive neighbor of C, i.e., u is adjacent to C and also to some i-novo-component $D \neq C$; or (2) u is not adjacent to C and is adjacent to some exclusive neighbor of C.

Claim 19. *If C is an i-novo-component at the beginning of round r, and there is more than one i-novo-component at that time, then there are at least k critical nodes for C in round r.*

Proof. Since the graph is k-vertex-connected, there are k internally-disjoint paths between C and other i-novo-components. Consider a collection of k such internally-disjoint paths, $P = \{p_1, \ldots, p_k\}$, for which their summed length is minimized. These paths have the following properties.

- If path $p_j \in P$ has length two, that is, $p_j = uxv$, where $u \in C$ and $v \in D$ for some i-novo-component $D \neq C$, then x is a common (non-exclusive) neighbor of C and D. Thus, x is a critical node for C.

- If path $p_j \in P$ has length three, that is, $p_j = uyzv$, where $u \in C$ and $v \in D$ for an i-novo-component $D \neq C$, we argue that y is an exclusive neighbor of C and z is not adjacent to C, thus z is a critical node for C: If y is not an exclusive neighbor of C, then it is a non-exclusive neighbor of C, i.e., there is an i-novo-component $D' \neq C$ and a node $v' \in D'$ such that y and v' are neighbors. Then path uyv' is shorter than p_j and internally-disjoint with all other paths p_m, $m \neq j$, contradicting the minimality of P; thus y must be an exclusive neighbor of C. If z is adjacent to C, i.e., it is adjacent to some $u' \in C$, then path $u'zv$ is shorter than p_j, and again we reach a contradiction as before; thus z cannot be adjacent to C.

- No path $p_j \in P$ has length greater than three. Suppose, for contradiction, that p_j has length greater than three, and let w be the last node in this path that is adjacent to C. If w is a non-exclusive neighbor of C, then it follows that we can replace p_j by a path of length two with internal node w. If w is an exclusive neighbor of C, and s is the next node in path p_j after w, then s must be adjacent to some other i-novo-component, because the set of sampled nodes is a dominating set and s is not adjacent to C. In this case, we can replace p_j by a path of length three with internal nodes w, s. So, in both cases we have a contradiction on the minimality of P.

It follows that each of the k internally-disjoint paths $p_j \in P$ has an internal node which is critical for C in round r, and this implies the claim. □

Next bound from below the drop in the number of j-novo-components in a round, for $j \leq i$, in terms of the number of i-novo components for which some critical node is sampled in the round. We start with an auxiliary claim, which shows some properties of the edges that are colored in the sub-steps of a step t.

Claim 20. *Suppose that in step t node u is sampled, and $\mathcal{C}$ is the set of all i-novo-components at the beginning of step t for which u is not adjacent to C and is adjacent to an exclusive neighbor of C.*

(a) For each edge uv that is colored with color-i in step t, node v is an exclusive neighbor of a distinct i-novo-component $C \in \mathcal{C}$ at the beginning of step t.

(b) For each i-novo-component $C \in \mathcal{C}$, some edge uv is colored with a color from $\{$color-$1, \ldots,$ color-$i\}$ in step t, where v is an exclusive neighbor of C at the beginning of step t.

We remark that (a) implies that for each i-novo-component C, at most one edge uv is colored with color-i in step t, where v is adjacent to C. Also (b) implies that the total number of edges uv that are colored with some color from $\{$color-$1, \ldots,$ color-$i\}$ is at least $|\mathcal{C}|$.

Proof. We show (a) first. Suppose edge uv is colored with color-i in step t. From of the edge-coloring procedure, v must be an exclusive neighbor of some i-novo-component C at the beginning of step t (Rule 2), and u is not adjacent to C (Rule 3). Thus, $C \in \mathcal{C}$. Further, if another edge uv', with $v' \neq v$, gets colored in step t with color-i, then v' cannot be an exclusive neighbor of C, otherwise Rule 4 would be violated, as it implies that v and v' are not adjacent to the same i-novo-component at the beginning of step t.

Next we prove (b). For an i-novo-component $C \in \mathcal{C}$, let v be an exclusive neighbor of C at the beginning of step t which is adjacent to u (there must be at least one such v, as $C \in \mathcal{C}$). Suppose that edge uv is not

colored with a color from $\{\text{color-}1, \ldots, \text{color-}i\}$ in step t. Edge uv is white before step t, as neither of u, v is sampled by that time, and is still white after the first i sub-steps by the above assumption. Then uv fulfilled all requirements for becoming color-i in sub-step i (Rules 1–3): edge uv was white before sub-step i; v was an exclusive neighbor of i-novo-component C at the beginning of step t; and u is not adjacent to C. Since uv is still white after sub-step i, it must be that, by Rule 4, some edge uv' was colored in sub-step i, where v' was another exclusive neighbor of C at the beginning of step t. □

The next claim shows that if some critical node for an i-novo-component C is sampled in round r, but C does not merge with other i-novo-components in this round, then some edge to an exclusive neighbor of C is colored with color-j, where $j < i$.

Claim 21. *Suppose that C is an i-novo-component at the beginning of round r, and some critical node for C is sampled in this round. Suppose also that C does not merge with other i-novo-components in round r. Then there is a step t of round r, in which a node u is sampled, and some edge uv is colored with color-j, where $j \leq i - 1$ and v is an exclusive neighbor of C at the beginning of round r.*

Proof. Let t be the earliest step of round r in which some critical node for C is sampled, and let u be the node sampled in this step. From the assumption that C does not merge with another i-novo-component in round r, it follows that u cannot be a non-exclusive neighbor of C at the beginning of round r, as this would imply that u was also adjacent to some other i-novo-component $D \neq C$ at that time, and thus sampling u would create a black path between C and D.

Hence, from Definition 18, we know that u is not adjacent to C and is adjacent to an exclusive neighbor v of C at the beginning of round r. Node u is also adjacent to some i-novo-component $D \neq C$ at the beginning of round r, as u is not adjacent to C and the union of all i-novo-components is a dominating set. This implies that u is not adjacent to C', otherwise sampling u would create a black path between C' and D, and thus an i-novo-path between C and D.

We now argue that the exclusive neighbor v of C which is adjacent to u at the beginning of round r is also an exclusive neighbor of C' at the beginning of step t: First, v is not sampled before step t, otherwise v would belong to C', and u would be adjacent to C'. Thus v is adjacent to C'. Second, v cannot be a non-exclusive neighbor of C', otherwise some neighbor u' of v must be sampled before step t in round r, where u' is not adjacent to C', implying that u' is not adjacent to C, and thus u' must be a critical node for C for round r. But this contradicts assumption, that u is the earliest-sampled critical node for C in round r.

We have thus far established that u is not adjacent to i-novo-component C' and is adjacent to node v which is an exclusive neighbor of C' at the beginning of step t. From Claim 20(b) then it follows that some edge uv' is colored with color-j in step t, where $j \leq i$ and v' is an exclusive neighbor of C' at the beginning of step t. To complete the proof of the claim it suffice to show that $j \neq i$, and that v' is also an exclusive neighbor of C at the beginning of round r.

If edge uv' above was colored with color-i, then Claim 16 would imply that C' merges with other i-novo-components in step t (in particular, at least with the i-novo-components adjacent to u at the beginning of step t). This contradicts the assumption that C (and thus C') does not merge with other i-novo-components in round r. Thus uv' is colored with color-j for some $j \leq i - 1$.

Finally, we show that v' is an exclusive neighbor of C at the beginning of round r. Suppose otherwise, towards a contradiction. Then v' is not adjacent to C or is a non-exclusive neighbor of C at the beginning of round r. In either case, v' must be adjacent to some i-novo-component $D \neq C$ at the beginning of round r, where in case v' is not adjacent to C, this is true because the union of i-novo-components is a dominating set. But then v' is also adjacent to some i-novo-component $D' \supseteq D$ at the beginning of step t, contradicting the assumption that v' is an exclusive neighbor of C'. □

For $1 \leq i \leq \lambda$ and $r \geq 0$, let $X_{i,r}$ denote the number of i-novo-components after the first r rounds. Thus the drop in the number of i-novo-components in round $r \geq 1$ is $X_{i,r-1} - X_{i,r}$.

Claim 22. *For any $1 \leq i \leq \ell$ and $r \geq 1$, we have that $2 \cdot \sum_{j=1}^{i}(X_{j,r-1} - X_{j,r})$ is at least equal to the number of i-novo-components at the beginning of round r for which a critical node gets sampled in round r.*

Proof. Let a be the number of i-novo-components at the beginning of round r for which some critical node is sampled in round r. Among these, let $C_1, \ldots, C_b$ be the i-novo-components which do not merge with other i-novo-components in round r. Hence the remaining $a - b$ merge with at least one other i-novo-component.

From Claim 21, we have that for each i-novo-component C_s, $1 \leq s \leq b$, there is a step t_s, in which some node u_s is sampled and an edge $u_s v_s$ is colored with color-j_s, where $j_s \leq i-1$ and v_s is an exclusive neighbor of C_s at the beginning of round r.

Observe that nodes v_s, $1 \leq s \leq b$, are distinct as they are exclusive neighbors of different i-novo-components at the beginning of round r. On the other hand, nodes u_s, $1 \leq s \leq b$, are not necessarily distinct.

Fix a step t and some $j \in \{1, \ldots, \leq i-1\}$, and suppose that node u is sampled in step t. Let $I(t,j)$ be the set of all $s \in \{1, \ldots, b\}$ for which $t_s = t$ and $j_s = j$ (then $u_s = u$, as well); assume $I(t,j) \neq \emptyset$. Then all edge uv_s, $s \in I(t,j)$, are colored with color-j in step t. By applying Claim 20(a) for these edges, we obtain that each node v_s, $s \in I(t,j)$, is an exclusive neighbor of a *distinct* j-novo-component D_s at the beginning of step t, and u is not adjacent to D_s.

To recap, for each $s \in I(t,j)$, D_s is a distinct j-novo-components at the beginning of step t, the node u sampled in step t is adjacent to exclusive neighbor v_s of D_s and not adjacent to D_s, and edge uv_s is colored with color-j in step t. Claim 16 then implies that all D_s, $s \in I(t,j)$, merge into a single j-novo-component. In fact, the total number of j-novo-components that merge in step t is at least $|I(t,j)| + 1$ as u must be adjacent to at least one j-novo-component D at the beginning of step t, and D is distinct from any D_s, as D_s is not adjacent to u. Therefore, the drop in the number of j-novo-components in step t is at least $|I(t,j)|$.

Applying the above to all steps t of round r and all $j \leq i-1$, and summing the corresponding drops, we obtain $\sum_{j=1}^{i-1}(X_{j,r-1} - X_{j,r}) \geq b$.

Recall that among all i-novo-components at the beginning of round r for which a critical node is sampled in round r, $a - b$ of them merge with some other i-novo-component in round r. This implies that the drop in the number of i-novo-components in round r is $X_{i,r-1} - X_{i,r} \geq (a-b)/2$.

It follows that $\sum_{j=1}^{i}(X_{j,r-1} - X_{j,r}) \geq (a+b)/2 \geq a/2$. □

We will use Claims 19 and 22 to show next that the expected drop in a round of the total number of j-novo-components for all $j \leq i$ is bounded from below by a linear function in the expected number of i-novo-components at the beginning of the round.

Claim 23. *For any $1 \leq i \leq \lambda$ and $r \geq 1$, and for $\rho = \frac{e-1}{2e}$, we have*

$$\sum_{j=1}^{i} \mathbf{E}[X_{j,r-1} - X_{j,r}] \geq \rho \cdot (\mathbf{E}[X_{i,r-1}] - 1).$$

Proof. By Claim 19,every i-novo-component C at the beginning of round r has at least k critical nodes in round r, as long as there are more than one such component. The probability that a node gets sampled in a given round is $1/k$, thus the probability at least one of the critical nodes for C gets sampled in round r is at least $1 - (1-1/k)^k \geq 1 - 1/e = 2\rho$. Then, given the number $X_{i,r-1}$ of i-novo-components C at the beginning of round r, the expected number of i-novo-components C for which some critical node is sampled is at least $2\rho X_{i,r-1}$, if $X_{i,r-1} > 1$.

Moreover, from Claim 22, we have that $\sum_{j=1}^{i}(X_{j,r-1} - X_{j,r})$ is at least equal to half the number of i-novo-components C for which some critical node is sampled. It follows that if $X_{i,r-1} > 1$,

$$\mathbf{E}\left[\sum_{j=1}^{i}(X_{j,r-1} - X_{j,r}) \;\middle|\; X_{i,r-1}\right] \geq \rho X_{i,r-1}.$$

Then $\mathbf{E}\left[\sum_{j=1}^{i}(X_{j,r-1}-X_{j,r}) \mid X_{i,r-1}\right] \geq \rho\cdot(X_{i,r-1}-1)$, and this inequality holds also (trivially) when $X_{i,r-1}=1$. Taking now the unconditional expectation yields

$$\mathbf{E}\left[\sum_{j=1}^{i}(X_{j,r-1}-X_{j,r})\right] \geq \rho\cdot(\mathbf{E}[X_{i,r-1}]-1). \qquad \square$$

Using Claim 23 we establish an upper bound on the expected number of i-novo-components after r rounds. In the following, it is more convenient to work with random variables $Y_{i,r}=X_{i,r}-1$, rather than directly with $X_{i,r}$, and we let $y_{i,r}=\mathbf{E}[Y_{i,r}]$. Claim 23 then implies

$$\sum_{j=1}^{i}(y_{j,r-1}-y_{j,r}) \geq \rho y_{i,r-1}. \tag{1}$$

Claim 24. *For any $1\leq i\leq\lambda$ and $r\geq 0$, and for $\rho=\frac{e-1}{2e}$ as in Claim 23, we have*

$$y_{i,r} \leq n\left(1-\frac{\rho}{2}\right)^{r}\left(1+\frac{2}{\rho}\right)^{i-1}.$$

Proof. We prove the statement by induction on r. For $r=0$, we have $y_{i,r}\leq n$ and thus the claimed inequality clearly holds for all $i\in\{1,\ldots,\lambda\}$.

For the induction step, we assume that the inequality holds for $y_{i,r-1}$ for all $i\in\{1,\ldots,\lambda\}$, for some $r\geq 1$, and bound $y_{i,r}$. Solving the inequality in (1) for $y_{i,r}$ and using the trivial lower bound $y_{j,r}\geq 0$ for all $j\leq i-1$, gives

$$y_{i,r} \leq (1-\rho)y_{i,r-1}+\sum_{j=1}^{i-1}y_{j,r-1}.$$

Applying the induction hypothesis to all terms on the right-hand side, we obtain

$$\begin{aligned}
y_{i,r} &\leq n\left(1-\frac{\rho}{2}\right)^{r-1}\left[(1-\rho)\left(1+\frac{2}{\rho}\right)^{i-1}+\sum_{j=1}^{i-1}\left(1+\frac{2}{\rho}\right)^{j-1}\right] \\
&< n\left(1-\frac{\rho}{2}\right)^{r-1}\left[\left(1+\frac{2}{\rho}\right)^{i-1}\left(-\rho+\sum_{h=0}^{\infty}\left(1+\frac{2}{\rho}\right)^{-h}\right)\right] \\
&= n\left(1-\frac{\rho}{2}\right)^{r-1}\left[\left(1+\frac{2}{\rho}\right)^{i-1}\left(1-\frac{\rho}{2}\right)\right],
\end{aligned}$$

and thus the claim follows. $\square$

Using Claim 24 and Markov's inequality we bound the number of rounds before there is just a single λ-novo-component left.

Claim 25. *All λ-novo-components have merged into a single component after 16λ rounds, with probability at least $1-n/2^{\lambda}$.*

Proof. The probability there is more than one λ-novo-component after the first r rounds is $\Pr(X_{\lambda,r}>1)=\Pr(Y_{\lambda,r}>0)=\Pr(Y_{\lambda,r}\geq 1)\leq\mathbf{E}[Y_{\lambda,r}]/1$, by Markov's inequality. Also from Claim 24,

$$\mathbf{E}[Y_{\lambda,r}] \leq n\left(1-\frac{\rho}{2}\right)^{r}\left(1+\frac{2}{\rho}\right)^{\lambda}.$$

Thus, in order to have $\Pr(X_{\lambda,r} > 1) \leq n/2^\lambda$, it suffices that

$$n\left(1-\frac{\rho}{2}\right)^r\left(1+\frac{2}{\rho}\right)^\lambda \leq n/2^\lambda.$$

Solving for r and substituting $\rho = \frac{e-1}{2e}$, we obtain $r \geq \lambda \ln\left(\frac{\rho}{2\rho+4}\right)/\ln\left(1-\frac{\rho}{2}\right) \approx 15.6085 \cdot \lambda$. □

We now show that if there is just one i-novo-component after t steps, then the set S_t of nodes that have been sampled by that time is i-semi-connected, i.e., for any partition of S_t into two sets T and $S_t \setminus T$ with no edges between them, the two sets have at least i common neighbors. We will use the next simple claim.

Claim 26. *At any time, if A is the set of sampled nodes, and B, $A \setminus B$ is a partition of A such that B and $A \setminus B$ are not connected and a j-novo-path exists between them, where $1 \leq j \leq \ell$, then a shortest j-novo-path between B and $A \setminus B$ has length exactly 2.*

Proof. As j-novo-paths only consist of color-j, gray, and black edges, at least one of the endpoints of each edge in a j-novo-path has to be sampled, and therefore at least every second node on a j-novo-path has to be sampled. Consider the sequence of sampled nodes on the j-novo-path between B and $A \setminus B$. Since the path has one endpoint in B and one in $A \setminus B$, there must exist two consecutive sampled nodes in the above sequence such that one is in B and the other in $A \setminus B$. Because they are consecutive in the sequence, their distance in the path is at most 2. Since the sets are not connected, their distance has to be exactly 2. Any shortest j-novo-path connecting B and $A \setminus B$ therefore has to be of length 2. □

Claim 27. *If there is only one i-novo-component after t steps, then S_t is i-semi-connected.*

Proof. Suppose that there is only one i-novo-component after t steps. We show that for any partition of S_t into two sets T and $S_t \setminus T$ with no edges between them, there is a j-novo-path between T and $S_t \setminus T$ for each $1 \leq j \leq i$, such that all these paths have length 2 and are internally disjoint. This implies that S_t is i-semi-connected.

Since there is just a single i-novo-component after t steps, Claim 17 gives that there is also just a single j-novo-component, for any $j \leq i$. Hence, for any $j \leq i$, there must be at least one j-novo-path connecting T and $S_t \setminus T$, and from Claim 26, there is a shortest j-novo-path of length 2 connecting T and $S_t \setminus T$.

For each $j \leq i$, consider the earliest j-novo-path of length 2 created between T and $S_t \setminus T$ (if more than one such path was created at the same time, we choose an arbitrary one among them). Let $u_j w_j v_j$ denote that path, where $u_j \in T$, $v_j \in S_t \setminus T$, and $w_j \notin S_t$. We will show that these paths are internally disjoint, i.e., all nodes w_j, for $1 \leq j \leq i$, are distinct.

Fix some $j \leq i$, and consider the time in which j-novo-path $u_j w_j v_j$ was created. At this time one of the edges $u_j w_j$ or $v_j w_j$ becomes color-j while the other edge is gray or became color-j in an earlier step. Assume w.l.o.g. that $v_j w_j$ is the edge that becomes color-j.

We argue that there is no gray edge $w_j x$, where $x \in S_t \setminus T$: Suppose there is such a gray edge $w_j x$. Then edge $u_j w_j$ cannot be color-j, because then $u_j w_j x$ is a j-novo-path created before $u_j w_j v_j$. Thus edge $u_j w_j$ is gray. However, it must be the case that before $v_j w_j$ became color-j, node w_j was an exclusive neighbor of a j-novo-component, and since both $u_j w_j$ and $w_j x$ are gray, it follows that u_j and x belonged to the same j-novo-component. Thus before $v_j w_j$ became color-j there was already a j-novo-path between nodes $u_j \in T$ and $x \in S_t \setminus T$, and thus there was also a j-novo-path of length 2 between two nodes from these two sets, by Claim 26. This contradicts that $u_j w_j v_j$ was the earliest such j-novo-path.

We now show that path $u_j w_j v_j$ is internally disjoint from ℓ-novo-path $u_\ell w_\ell v_\ell$, for any $\ell < j$, i.e., $w_j \neq w_\ell$. Suppose that $w_j = w_\ell$, for some $\ell < j$. Since we have shown that there is no gray edge between w_j and some node from $S_t \setminus T$, it must be that edge $w_\ell v_\ell$ is color-ℓ. We argue that v_ℓ is sampled *after* v_j: Suppose, for contradiction, that v_ℓ is sampled before v_j. Then when v_j is sampled, v_ℓ and u_ℓ must be in

the same j-novo-component, otherwise w_j is adjacent to two distinct j-novo-components, preventing $v_j w_j$ from becoming color-j. Thus before $v_j w_j$ became color-j there was a j-novo-path between nodes $u_\ell \in T$ and $v_\ell \in S_t \setminus T$, and thus there was also a j-novo-path of length 2 between two nodes from these two sets, by Claim 26. This contradicts that $u_j w_j v_j$ was the earliest such j-novo-path. We conclude that v_ℓ was sampled *after* v_j. However, if v_ℓ was sampled after v_j, this means that the ℓ-novo-path created when edge $w_j v_\ell$ was colored, cannot be the earliest such path created between T and $S_t \setminus T$, because that path must have been created no later than the earliest j-novo-path, by Claim 17. We have thus established that $w_j \neq w_\ell$, for any $1 \leq \ell < j \leq i$, thus completing the proof of Claim 27. □

By Claim 25, the sampling procedure results in a single λ-novo-component after at most 16λ rounds, with probability at least $1 - n/2^\lambda$. In each round the sampling probability is $1/k$, thus the total sampling probability is at most $16\lambda/k$. Once there is just a single λ-novo-component, by Claim 27 we have that the set S_t of sampled nodes is λ-semi-connected. This concludes the proof of Lemma 14.

4 Proof of Propositions 4 and 5

In this section we prove the two statements from Section 1.2 that demonstrate the optimality of the bound in Theorem 1.

Proof of Proposition 4. The edge connectivity of G is at most k as it contains an edge-cut of size k, and thus its vertex connectivity is also at most k. On the other hand, it is easy to verify that the removal of any $k-1$ vertices does not disconnect G. Therefore G has vertex connectivity exactly k.

Let K denote the number of edges in the matching that survive after sampling (i.e., both their endpoint nodes are sampled). The expected value of K is $\mathbf{E}[K] = kp^2$, since each edge survives with probability p^2. If $K \neq 0$ then K is an upper bound on the edge connectivity and thus on the vertex connectivity of the sampled subgraph. If $K = 0$ then it is still possible for the vertex connectivity to be positive, if no nodes are sampled from the one clique and at least one is sampled from the other. Let N_i, for $i = 1, 2$, denote the number of nodes sampled in each of the two cliques respectively, and let Z_i be the indicator random variable with $Z_i = 1$ if $N_i = 0$ and $Z_i = 0$ otherwise. Then $\mathbf{E}[N_i] = pn$, and $\mathbf{E}[Z_i] = \Pr(N_i = 0) = (1-p)^n$. From the discussion above it follows that the vertex expansion of the sampled subgraph is at most $K + Z_2 N_1 + Z_1 N_2$, and thus the expected vertex expansion is at most

$$\begin{aligned}\mathbf{E}[K + Z_2 N_1 + Z_1 N_2] &= \mathbf{E}[K] + 2\,\mathbf{E}[Z_2 N_1] = \mathbf{E}[K] + 2\,\mathbf{E}[Z_2] \cdot \mathbf{E}[N_1] \\ &= kp^2 + 2np(1-p)^n \leq kp^2 + 2npe^{-np}.\end{aligned}$$

If $p \geq 2\ln n/n$, then the second term in the last line above is $kp^2 \cdot (2n/kp)e^{-np} \leq kp^2 \cdot (1/k\ln n) = o(kp^2)$; thus the expected vertex connectivity is at most $kp^2 + o(kp^2)$.

For the probability that the sampled subgraph is disconnected, we first observe that if $p = O(1/n)$ then the subgraph is empty (and thus by convention disconnected) with constant probability. Thus, below we assume that $p \geq 2/n$. The probability that the sampled subgraph is disconnected is bounded from below by

$$\begin{aligned}\Pr(K = 0 \wedge N_1 \neq 0 \wedge N_2 \neq 0) &\geq 1 - (\Pr(K \neq 0) + \Pr(N_1 = 0) + \Pr(N_2 = 0)) \\ &= \Pr(K = 0) - 2\Pr(N_1 = 0) \\ &= (1-p^2)^k - 2(1-p)^n.\end{aligned}$$

The second term in the last line is at most $(1-p^2)^k/2$, as

$$\frac{2(1-p)^n}{(1-p^2)^k} \leq \frac{2(1-p)^n}{(1-p^2)^n} = \frac{2}{(1+p)^n} \leq \frac{2}{(1+2/n)^n} \leq \frac{1}{2}.$$

It follows that the probability of the sampled subgraph to be disconnected is at least $(1-p^2)^k/2$, and this is at least $1/n^{o(1)}$ if $p = o(\sqrt{\log(n)/k})$. □

Proof Sketch of Proposition 5. Since $p = \omega(1/n)$, we have with probability $1 - o(1)$ that at least one node gets sampled from the first $n/3k$ cliques, and at least one gets sampled from the last $n/3k$ cliques. The probability that no edge survives in the cut between two given consecutive cliques is $(1-p^2)^k = e^{-o(\log(n/k))} = \omega(k/n)$, as $p = o(\sqrt{\log(n/k)/k})$. Thus, the probability that at least one of the cuts between the middle $n/3k$ cliques gets disconnected is at least

$$1 - (1 - \omega(k/n))^{k/6k} = 1 - o(1),$$

where for this computation we just considered every second cut, i.e., $n/6k$ cuts in total, and used the fact that these cuts are vertex-disjoint. Combining the above yields the claim. □

5 Discussion

In this paper we show two main results: (1) When independently sampling *vertices* of a k-vertex-connected n-node graph with probability $p = \Omega(\sqrt{\log(n)/k})$, the sampled subgraph has a vertex connectivity of $\Omega(kp^2)$, with high probability; and (2) When independently sampling *edges* of a k-vertex-connected n-node graph with probability $p = \Omega(\log(n)/k)$, the sampled subgraph has a vertex connectivity of $\Omega(kp)$, with high probability. The core technical part, for both results, is to prove that vertex sampling with probability $p = \Omega(\sqrt{\log(n)/k})$ yields a subgraph that is (just) connected with sufficiently high probability. This is achieved by considering sampling as a gradual random process, and carefully analyzing the growth of the (novo-)connected components, using the novel notions of semi-connectivity and novo-connectivity.

The constant factors in our results are much smaller than 1; it would be interesting to identify the correct constants. Most importantly, we leave open whether the remaining vertex connectivity under vertex sampling is in fact at least $kp^2(1-\epsilon)$, for an arbitrary small $\epsilon > 0$, assuming kp^2 is large enough, e.g., $kp^2 = \Omega(\log n/\operatorname{poly}(\epsilon))$. In particular, for a sampling probability of $p = 1 - o(1)$, or equivalently a sub-constant deletion probability, this would imply a remaining connectivity of $k - o(k)$ instead of just $O(k)$. The same question can also be asked for the remaining vertex connectivity under *edge* sampling.

Our results show only *lower* bounds on the remaining vertex connectivity. There are k-vertex-connected graphs which, under the same sampling processes, would retain a much higher vertex connectivity, e.g., up to kp when sampling vertices, and up to k when sampling edges. It would be interesting to see if one can tightly characterize the (e.g., expected) remaining vertex connectivity under sampling of a given graph as a simple and natural function of it. Alternatively, is there a variant of these random sampling processes, which in effect sparsifies the graph, but for which we can tightly characterize the remaining vertex connectivity?

Finally, as stated in Corollary 6, our result implies that any graph can be partitioned into $\Omega(k/\log^2 n)$ vertex-disjoint connected dominating sets. While this is an improvement over the best previously known lower bound of $\Omega(k/\log^5 n)$, a logarithmic gap still remains compared to the upper bound of $O(k/\log n)$ for the number of vertex-disjoint connected dominating sets that is known for some graphs. Closing this gap is an intriguing open question for further research.

References

[1] N. Alon. A note on network reliability. In *Discrete Probability and Algorithms*, pages 11–14. Springer, 1995.

[2] A. A. Benczúr and D. R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *Proc. 28th ACM Symposium on Theory of Computing (STOC)*, pages 47–55, 1996.

[3] B. Bollobás. *Random graphs*. Springer, 1998.

[4] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press, 2006.

[5] K. Censor-Hillel, M. Ghaffari, and F. Kuhn. Distributed connectivity decomposition. In *Proc. 32nd ACM Symposium on Principles of Distributed Computing (PODC)*, pages 156–165, 2014.

[6] K. Censor-Hillel, M. Ghaffari, and F. Kuhn. A new perspective on vertex connectivity. In *Proc. 25th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 546–561, 2014.

[7] A. Kanevsky. On the number of minimum size separating vertex sets in a graph and how to find all of them. In *Proc. 1st ACM-SIAM Symposium on Discrete Algorithm (SODA)*, pages 411–421, 1990.

[8] D. R. Karger. Global min-cuts in $\mathcal{RNC}$, and other ramifications of a simple mincut algorithm. In *Proc. 4th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 21–30, 1993.

[9] D. R. Karger. Random sampling in cut, flow, and network design problems. In *Proc. 26th ACM Symposium on Theory of Computing (STOC)*, pages 648–657, 1994.

[10] D. R. Karger. Using randomized sparsification to approximate minimum cuts. In *Proc. 5th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 424–432, 1994.

[11] D. R. Karger. A randomized fully polynomial time approximation scheme for the all terminal network reliability problem. In *Proc. 27th ACM Symposium on Theory of Computing (STOC)*, pages 11–17, 1995.

[12] D. R. Karger and M. S. Levine. Finding maximum flows in undirected graphs seems easier than bipartite matching. In *Proc. 30th ACM Symposium on Theory of Computing (STOC)*, pages 69–78, 1998.

[13] M. V. Lomonosov and V. P. Polesskii. Lower bound of network reliability. *Problems of Information Transmission*, 8:118–123, 1972.

[14] C. S. J. A. Nash-Williams. Edge-disjoint spanning trees of finite graphs. *Journal of the London Mathematical Society*, 36(1):445–450, 1961.

[15] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. 36th ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2004.

[16] W. T. Tutte. On the problem of decomposing a graph into n connected factors. *Journal of the London Mathematical Society*, 36(1):221–230, 1961.

[17] A. Zehavi and A. Itai. Three tree-paths. *Journal of Graph Theory*, 13(2):175–188, 1989.

Appendix: Novo-Connectivity Implies Simple Novo-Paths

The next claim shows that two nodes are i-novo-connected iff a *simple* i-novo-path between them exists. Interestingly, this very natural property of novo-connectivity is not trivial to show.

Lemma 28. *If there is a non-simple i-novo-path between two nodes, then there is also a simple i-novo-path between them.*

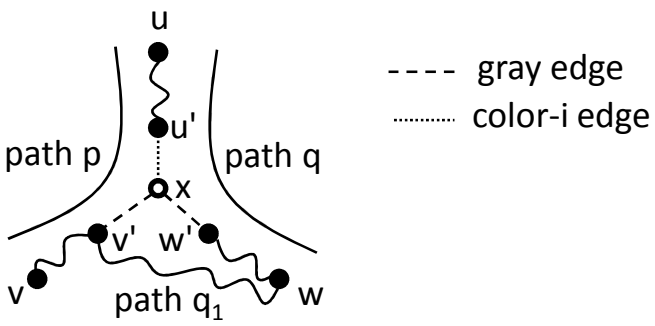

Figure 1: Illustrating the proof of Lemma 28

Proof. We will show that the following transitivity property holds: if there is a *simple* i-novo-path between nodes u and v, and between nodes u and w, then there is a *simple* i-novo-path between v and w.

From this, the main claim follows easily: Suppose there is a non-simple i-novo-path $u_0u_1 \dots u_\ell$ between nodes u_0 and u_ℓ. Divide this path into subpaths $u_0 \dots u_{i_1}$, $u_{i_1} \dots u_{i_2}, \dots$ such that the endpoints of each subpath are sampled nodes, while the internal nodes are not. It is immediate from the definition of novo-paths (Definition 15) that each of the subpaths has length 1 or 2 and is a simple i-novo-path. We can now repeatedly apply the transitivity property above to conclude that a simple i-novo-path exists between u_0 and u_ℓ.

Next we prove the transitivity property. Suppose, for the sake of contradiction, that the property is violated at some point, and let t be the earliest step when this happens. That is, at some point during step t, there is some i and nodes u, v, w such that there is a simple i-novo-path between u and v, and between u and w, but there is no simple i-novo-path between v and w. Recall that before the first step $t = 1$, there are no color-i edges, so at that time any i-novo-path consists of only black edges, and thus the transitivity property clearly holds.

Let p be a simple i-novo-path between v and u, and q a simple i-novo-path between w and u (see Figure 1). Let x be the first node where the two paths intersect when going from w towards u on path q. We define r to be the concatenation of the subpath of p connecting v and x and of the subpath of q connecting x and w. From our choice of x, it follows that r is also a *simple* path, between v and w, as x is the only node of r that is in the intersection of the simple paths p and q. Further, note that x cannot be a sampled node because in that case r is an i-novo-path connecting nodes v and w. Hence, in particular, $x \notin \{u, v, w\}$. Let v' and w' be the neighbors of x in path r towards v and w, respectively, and let u' be the neighbor of x in p towards u. Notice it is possible that $u = u'$, $v = v'$, or $w = w'$. We also observe that both edges xv' and xw' must be gray, because if at least one of them is color-i then r is an i-novo-path.

We have thus established that node x is not sampled and both edges xv' and xw' are gray. Since xv' and xu' are consecutive edges in i-novo-path p and x is not sampled, it follows that xu' must be color-i. Consider the step $t' \leq t$ at which this edge changed from white to color-i, when u' was sampled. It must be the case that before step t', and thus before step t, x was an exclusive neighbor to a single i-novo-component, by our coloring rules. We stress here that at any point before step t, the transitivity property of simple i-novo-paths holds because of the minimality of t, thus at any point before step t there is a simple i-novo-path between any two i-novo-connected nodes, as we argued at the beginning. Since xv' and xw' are both gray and since no edge *becomes* gray at any step, these edges were also gray before step t', which implies that v' and w' were in the same i-novo-component before step t.

We now argue that at least one of nodes v and w is also in the same i-novo-component as v' and w'

before step t: The subpath of r between v and v' and the subpath between w and w' are both simple i-novo-paths and they do not intersect. We also have that in step t, as in any step, only edges incident to the node sampled in that step (if one is indeed sampled) may change color. Since the subpaths above do not share a common node, at least one of them does not change in step t. This implies that at least one of v and w is in the same i-novo-component as v' and w' before step t, as desired. In case v is in the same i-novo-component as v' and w' before t, then there is a simple i-novo-path between v and w' before t; and if w is in the same i-novo-component as v' and w' before t, then a simple i-novo-path exists between w and v' before t.

Therefore, we have established that there are nodes a, b, c such that there is a simple i-novo-path between a and b, and between a and c, but not between b and c, and moreover the simple i-novo-path between a and b exists also before step t. (We saw above that these conditions are met for $(a, b, c) = (w', v, w)$ or (v', w, v).)

Among all node triples a, b, c satisfying the above conditions, we consider one for which the length of the shortest simple i-novo-path between a and c is *minimal*. Similar to the analysis before for u, v, w, let p' and q' be the simple i-novo-paths from a to b and c, respectively (so, p' exists also before step t, and q' is of minimal length). Let x' be the first node in the intersection of p' and q' when going from b towards a on path q', and let r' denote the concatenation of the subpaths of p' and q' connecting x' with a and b, respectively. As before, x' cannot be sampled, as otherwise r' is a simple i-novo-path. Defining a' and b' in a similar manner as v' and w', by the same argument as before we get that a' and b' are in the same i-novo-component before step t. Observe now that b is also in that i-novo-component, because b and b' are connected by an i-novo-path before step t, namely the subpath of p' between b and b' (this is where we use the assumption that p' exists before t).

Therefore, we have established that there is a simple i-novo-path between b and c' that exists also before step t, and there is a simple i-novo-path between c' and c which is a proper subpath of q', and is thus shorter than q'. This contradicts, the optimality condition based on which nodes a, b, c were selected, as nodes c', b, c would be a better choice. □